Improving Statistical Machine Translation with Target-Side Dependency Syntax(Abstract_要旨)

AUTHOR(S):

John, Walter Richardson

CITATION:

John, Walter Richardson. Improving Statistical Machine Translation with Target-Side Dependency Syntax. 京都大学, 2016, 博士(情報学)

ISSUE DATE:

2016-09-23

URL:

https://doi.org/10.14989/doctor.k20022

（ 続紙 1 ）

京都大学	博士（情報学）	氏名	John Walter Richardson
論文題目	Improving Statistical Machine Translation with Target-Side Dependency Syntax （目的言語側の依存構文による統計的機械翻訳の改善）		

（論文内容の要旨）

Machine Translation (MT) is the application of Natural Language Processing that focuses on the automatic translation between languages. Automatic translation is a widely used technology with wide-ranging applications, however translation can be challenging for language pairs with widely different grammatical structures, such as English and Japanese. Syntax-based MT is a translation paradigm based on the principle of generalizing language with grammatical rules that can be employed to improve translation between distant language pairs, as the additional layer of abstraction enables the design of more robust and flexible translation rules.

The majority of previous approaches to syntax-based MT have employed only source-side grammar, known as tree-to-string MT, and previous studies on exploiting target-side syntax, i.e. tree-to-tree MT, have not shown promising results. The aim of this thesis is to analyze in detail the effectiveness of target-side syntax in the modern world of machine translation. The thesis consists of 8 chapters as described below.

Chapter 1 begins with an overview of machine translation, outlining the major paradigms and methods of evaluation. The place of dependency tree-to-tree translation in modern MT research is described.

Chapter 2 outlines the case study of a state-of-the-art dependency tree-to-tree system, KyotoEBMT, which has been developed as a core component of this research on syntax-based MT. In this chapter, the design and extraction of dependency tree-to-tree translation rules is discussed and the proposed system is able to outperform the state of the art phrase-based SMT framework Moses by 1-3 BLEU points. Analysis of translation results gives empirical evidence of the advantages and disadvantages of syntax-based approaches and provides a starting point for the main investigation.

The thesis proceeds to analyze two major aspects of translation where target-side syntax can be effective: word order and translation fluency.

In Chapter 3, a preliminary exploration of fluency improvement with target-side syntax is presented. A simple tree-based language model for post-editing function words is proposed and shown to be effective for Japanese-English patent translation, giving a significantly positive improvement as judged by human raters.

Chapter 4 extends this work to a fully-fledged dependency tree language model. The idea of generalized dependency tree context with decomposed partial dependency trees, called t-treelets, is introduced. Experimental results show that translation fluency is improved significantly for morphologically rich languages.

Chapter 5 describes an improved language model exploiting source-side parses mapped to

the translation using word alignment. The source-side information is also included to construct a bilingual language model. This approach is able to achieve gains in translation quality for 20 language pairs as judged by human experts.

In Chapter 6, target-side syntax is applied to the task of reordering. In particular, the extraction of flexible dependency tree-to-tree translation rules is considered. These are shown to be effective at improving word order in translations between distant language pairs, achieving up to a 2 BLEU point improvement over a state-of-the-art tree-to-tree MT baseline.

Chapter 7 presents an extended exploration of potential avenues for future work. While this thesis concentrates on statistical syntax-based approaches, the field has recently seen a surge in interest in translation methods based on neural networks. This chapter considers how ideas can be shared between these new approaches and the proposed syntax-based models. The initial experiments presented are already able to show an improvement of 2 BLEU points for reranking and 1 BLEU for reordering.

The thesis concludes in Chapter 8 with a summary of the potential impact and future directions for the work presented.

注）論文内容の要旨と論文審査の結果の要旨は１頁を３８字×３６行で作成し，合わせて，３，０００字を標準とすること．
論文内容の要旨を英語で記入する場合は，４００～１，１００words で作成し
審査結果の要旨は日本語５００～２，０００字程度で作成すること．

（続紙 ２ ）

（論文審査の結果の要旨）

本論文は、従来の機械翻訳システムにおいて最も深刻な問題である翻訳文の語順と流暢さの改善を目的として、目的言語側の依存構文を用いる方法を提案するものである。具体的には依存構文言語モデルと柔軟な語順入れ替えモデルを提案した。得られた主要な成果は以下の通りである。

１．従来の構文に基づく機械翻訳システムは、主に原言語側のみの文構造を利用するものであり、言語構造が大きく異なる言語対の翻訳では語順の誤りが多いという問題があった。本論文では目的言語側の依存構文を考慮することにより、修飾要素の柔軟な挿入を可能とする翻訳ルールの生成方法を考案した。これによって翻訳語順の精度を向上させることに成功した。

２.従来の機械翻訳システムでは単語 n-gram による言語モデルが主に利用されてきたが、このモデルは数単語程度の局部的な情報しか考慮することができず、遠い位置にある単語の影響を考慮することができなかった。本研究では、文脈部分木に基づいた広い文脈を考慮する依存構文言語モデルを考案し、これを用いて翻訳の後編集を行う枠組みを提案した。本枠組みでは、目的言語の構文解析器がない場合でも原言語側の構文情報を目的言語側に射影して利用することが可能である。英語を原言語とし２０カ国語を目的言語とする実験を行い、すべての言語対において有意な改善が見られることを確認した。

３．これまでの機械翻訳の分野では統計的機械翻訳が広く研究されてきたが、近年ではニューラルネットワークを用いる機械翻訳システムに長足の進展が見られる。そこで、ニューラルネットワークで計算される素性を用いて、依存構文を考慮する機械翻訳の結果をリランキングすることにより、翻訳精度を大幅に改善できることを示し、両手法を統合することに成功した。

４．本論文で考案した目的言語の依存構文を利用する方式を京都大学用例翻訳システム、KyotoEBMT として実装した。このシステムは 2014 年、2015 年の Workshop on Asian Translation で上位の成績をおさめた。2014 年からオープンソースで公開し、広く利用可能とした。

よって、本論文は博士（情報学）の学位論文として価値あるものと認める。また、平成２８年８月３１日に実施した論文内容とそれに関連した試問の結果合格と認めた。

注)論文審査の結果の要旨の結句には、学位論文の審査についての認定を明記すること。
更に、試問の結果の要旨（例えば「平成　年　月　日論文内容とそれに関連した
口頭試問を行った結果合格と認めた。」）を付け加えること。

Ｗｅｂでの即日公開を希望しない場合は、以下に公開可能とする日付を記入すること。
要旨公開可能日：　　年　　月　　日以降